greatlearning
Power Ahead

Business Presentation

Presenter : Esteban Ordenes

Contents

1. Define the problem and perform an Exploratory Data Analysis
2. Illustration of the insights based on EDA
3. Applying K-means clustering algorithms
4. Applying Hierarchical clustering
5. Compare cluster K-means clusters and Hierarchical clusters
6. Actionable Insights & Recommendations

Business Problem Overview and Solution Approach

- AllLifeBank Customer Segmentation

- Core business idea
  - AllLife Bank wants to focus on its credit card customer base in the next financial year. They have been advised by their marketing research team, that the penetration in the market can be improved. Based on this input, the Marketing team proposes to run personalized campaigns to target new customers as well as upsell to existing customers. Another insight from the market research was that the customers perceive the support services of the back poorly. Based on this, the Operations team wants to upgrade the service delivery model, to ensure that customer's queries are resolved faster.
- Problem to tackle
  - Identify different segments in the existing customer based on their spending patterns as well as past interaction with the bank.

Data Overview

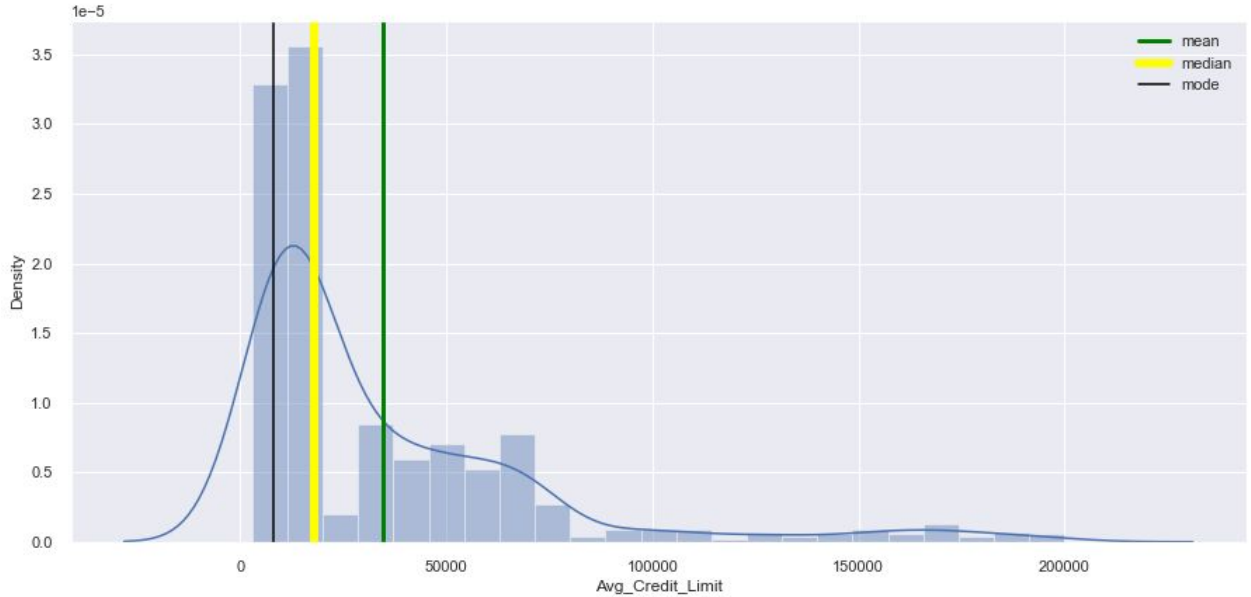
- **Data Dictionary**

Data is of various customers of a bank with their credit limit, the total number of credit cards the customer has, and different channels through which customer has contacted the bank for any queries, different channels include visiting the bank, online and through a call center.

- Sl_No : Serial Number
- Customer Key : Customer unique Key
- Avg_Credit_Limit : Average Credit Limit
- Total_Credit_Cards : Number of Credit Cards they have
- Total_visits_bank : Number of times they visited the bank
- Total_visits_online : Number of times they visited the website online
- Total_calls_made : Number of times they made a call

EDA - Feature : Avg_Credit_Limit

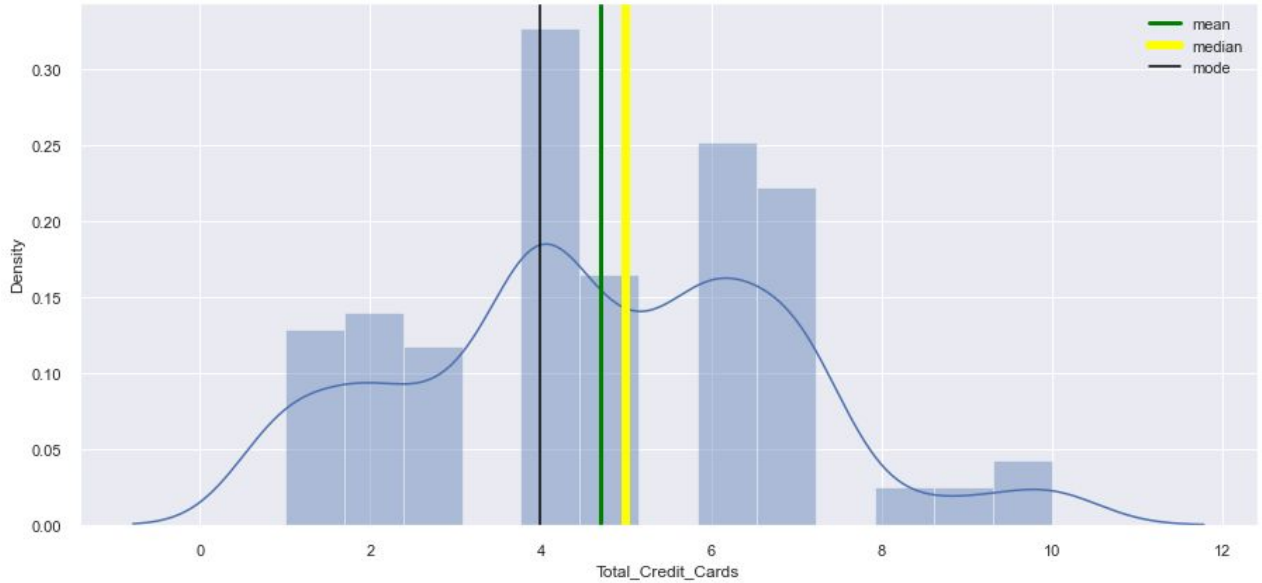
Mean:34878.274268104775
Median:18000.0
Mode:8000



- Avg_Credit_Limit feature is right-skewed.
- There are many outliers to the right of the curve

EDA - Feature : Total_Credit_Cards

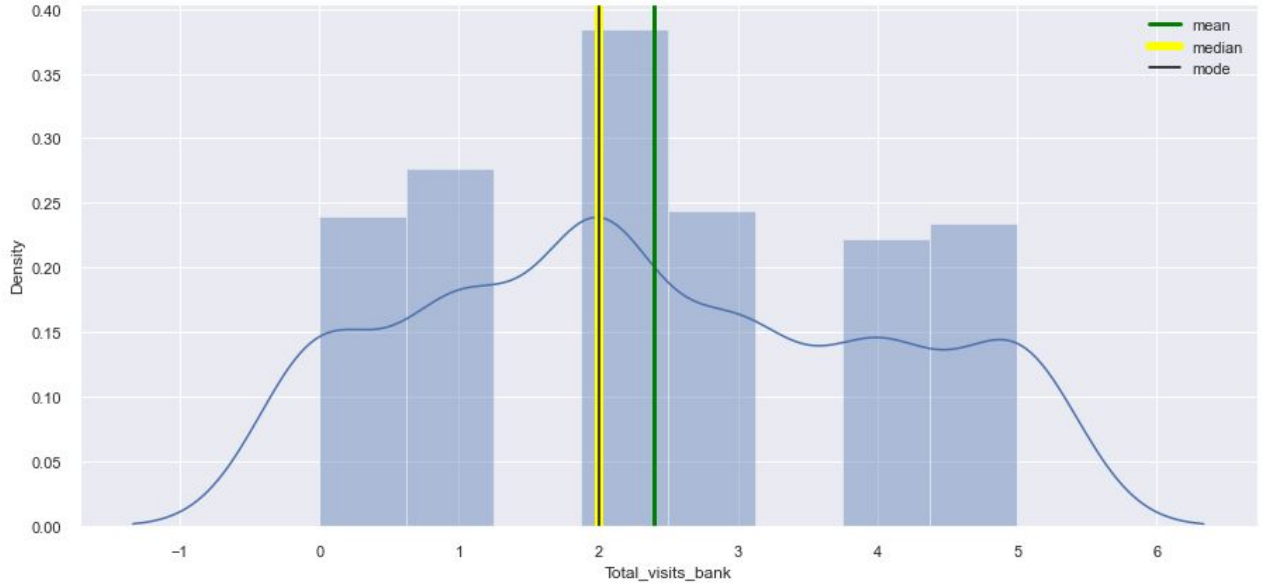
Mean:4.708782742681048
Median:5.0
Mode:4



- Total_Credit_Cards feature seems fairly normal distributed.
- There are no outliers in the distribution.

EDA - Feature : Total_visits_bank

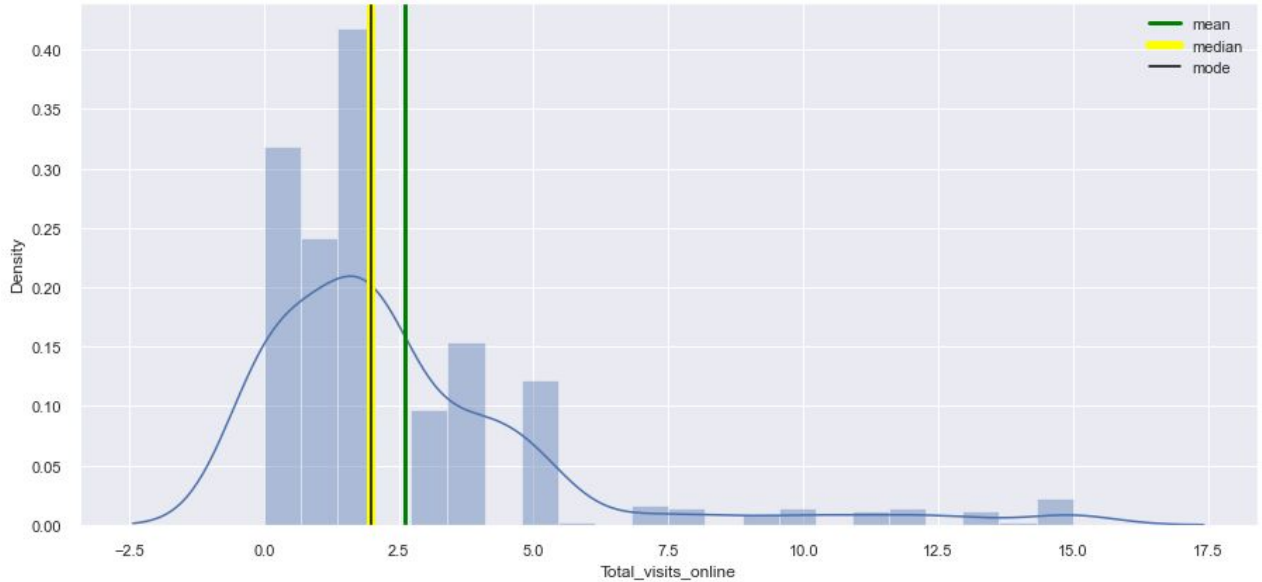
Mean:2.3975346687211094
Median:2.0
Mode:2



- Total_visits_bank seems fairly normal distributed.
- There are no outliers in the distribution.

EDA - Feature : Total_visits_online

Mean:2.6240369799691834
Median:2.0
Mode:2

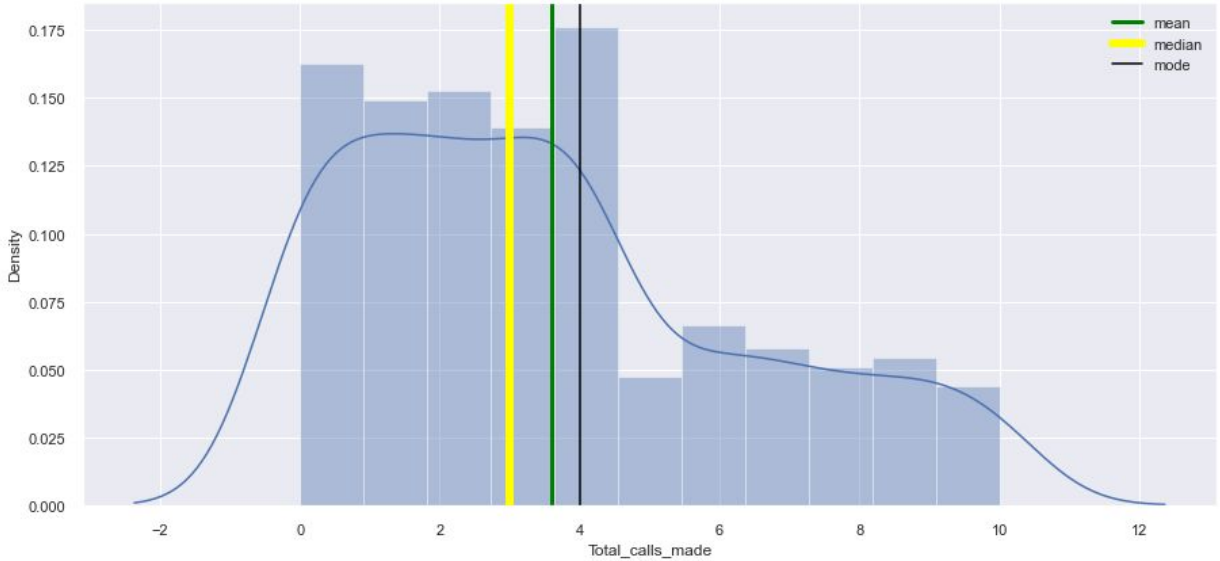


- Total_visits_online is right-skewed.
- There are many outliers to the right of the curve which may explain its skewness.

EDA - Feature : Total_calls_made

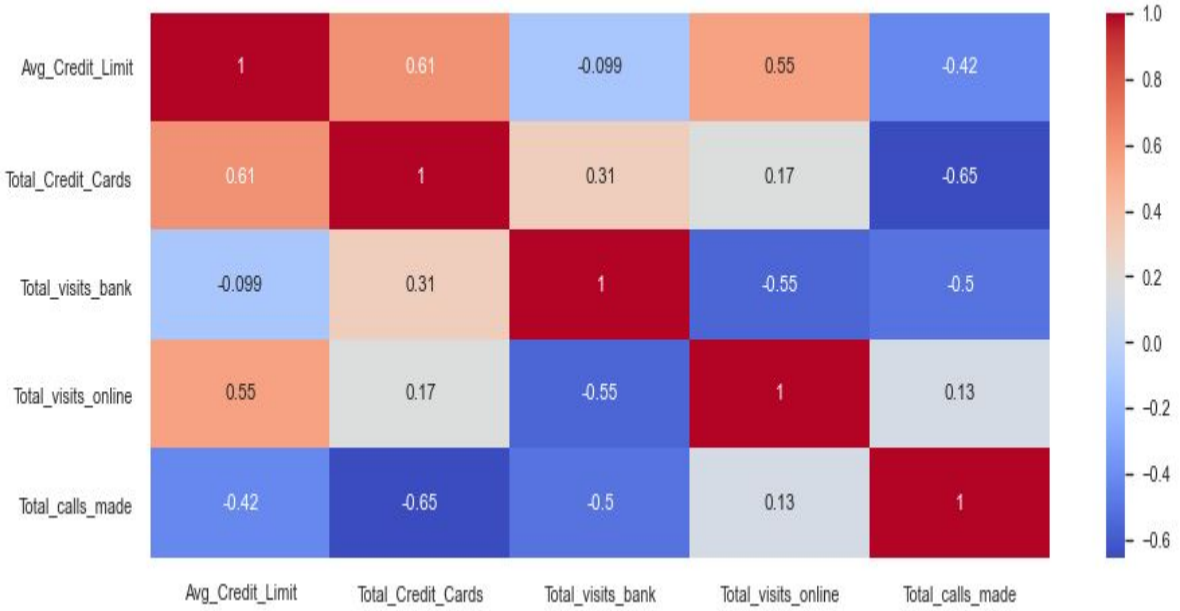
Mean:3.5901386748844377
Median:3.0
Mode:4

- Total_calls_made is right-skewed.
- There are no outliers.



EDA - **Bivariate Analysis**

- The pairs that have a positive/high (>0.5) correlation are:
  - Total_Credit_Cards/Avg_Credit_Limit
  - Total_visits_online/Avg_Credit_Limit
- The pairs that have a negative/high (<-0.5) correlation are:
  - Total_Credit_Cards/Total_calls_made
  - Total_visits_bank/Total_visits_online
  - Total_visits_bank/Total_calls_made

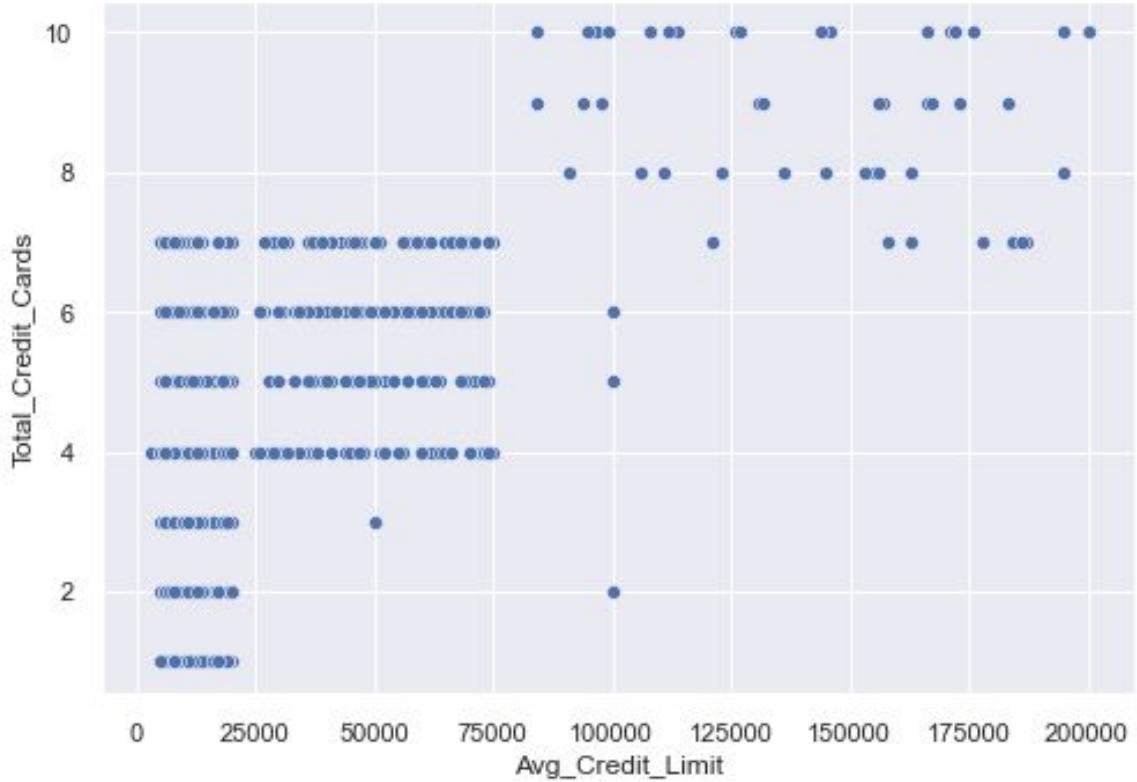


EDA - **Bivariate Analysis**

- There is no clear normally distributed feature
- Avg_Credit_limit and Total_visits_online are right-skewed
- Total_Credit_Cards and Total_calls_made seem multi-modal

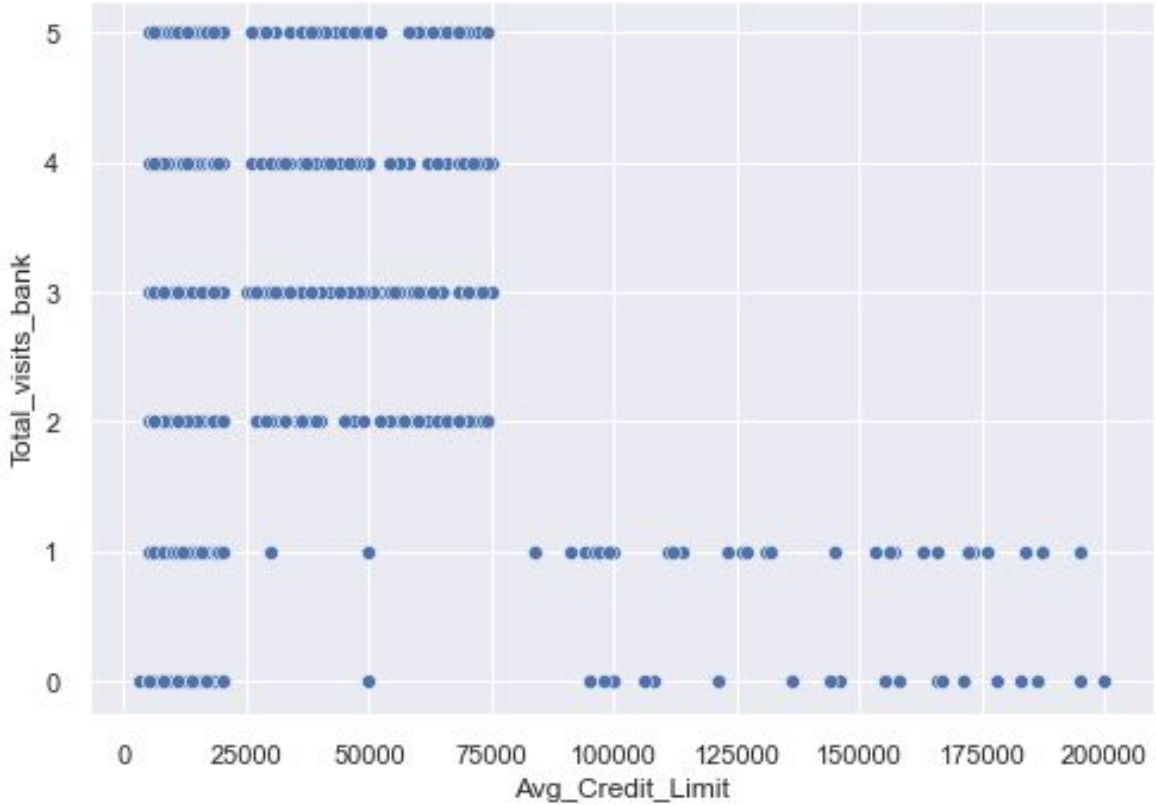
EDA - **Avg_Credit_Limit vs Total_Credit_Cards**

- It looks like those customers that have 7 or less Total_credit_cards have an Avg_Credit_limit of 75000 or less.
- As opposed to those that have a Total_Credit_Cards of 8 and above have an Avg_Credit_Limit above 75000.



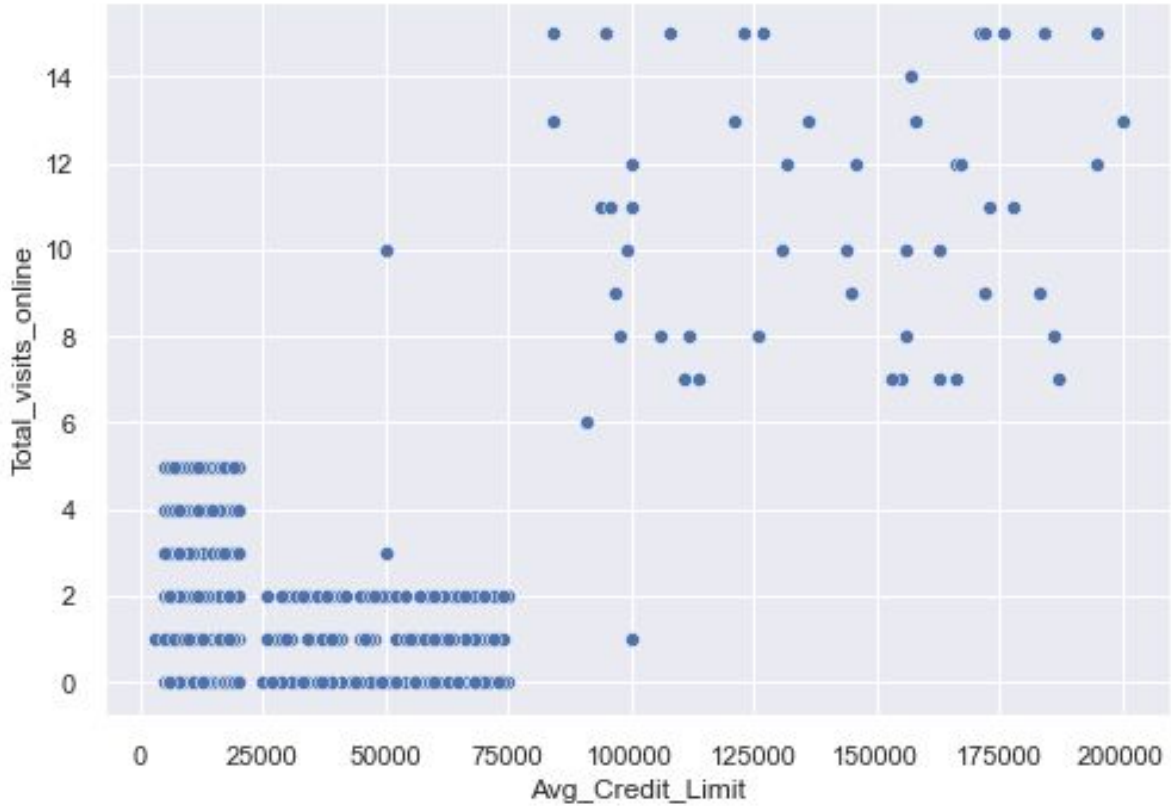
EDA - **Avg_Credit_Limit vs Total_visits_bank**

- It looks like those customers that have 1 or 0 Total_visits_bank have an Avg_Credit_limit either 25000 or below OR 85000 and above.
- Those customers that have Total_visits_bank 2 and above have an Avg_Credit_Limit of 75000 or less.



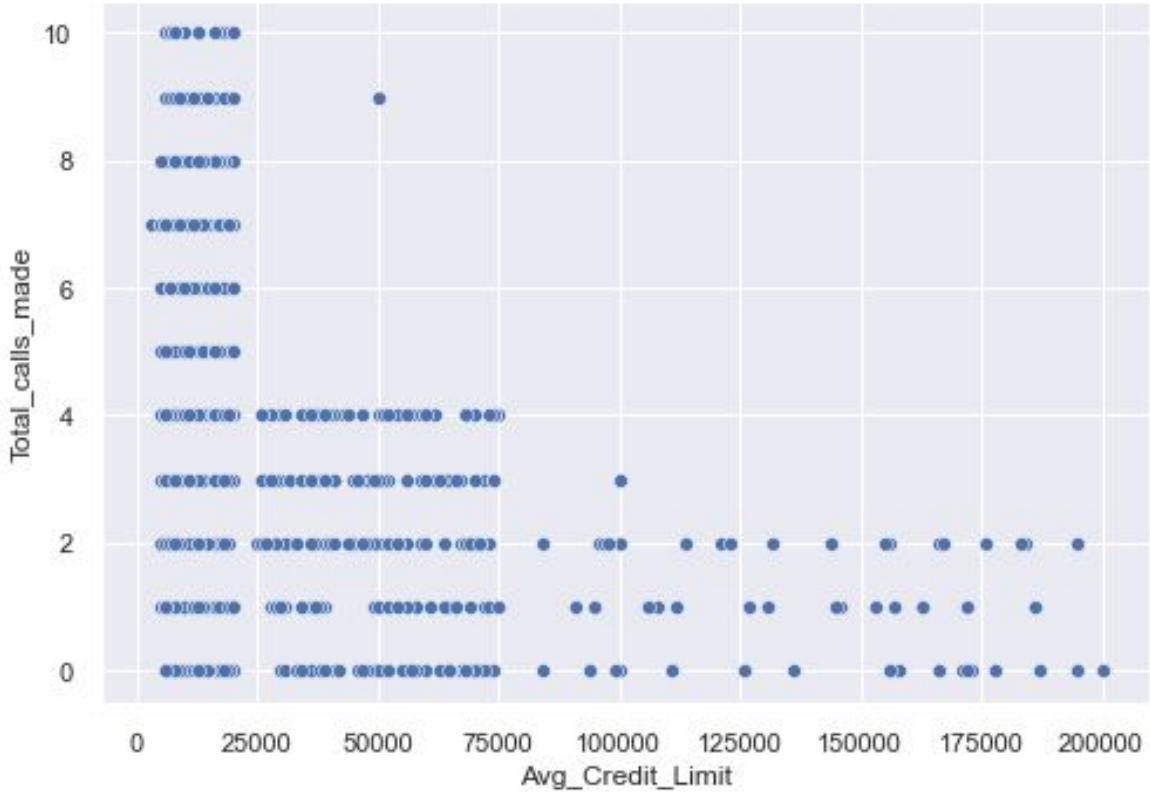
EDA - **Avg_Credit_Limit vs Total_visits_online**

- It looks like those customers that have 6 or less Total_visits_online have an Avg_Credit_limit below 75000.
- While those customers that have Total_visits_online above 6, have an Avg_Credit_Limit above 75000.



EDA - **Avg_Credit_Limit vs Total_calls_made**

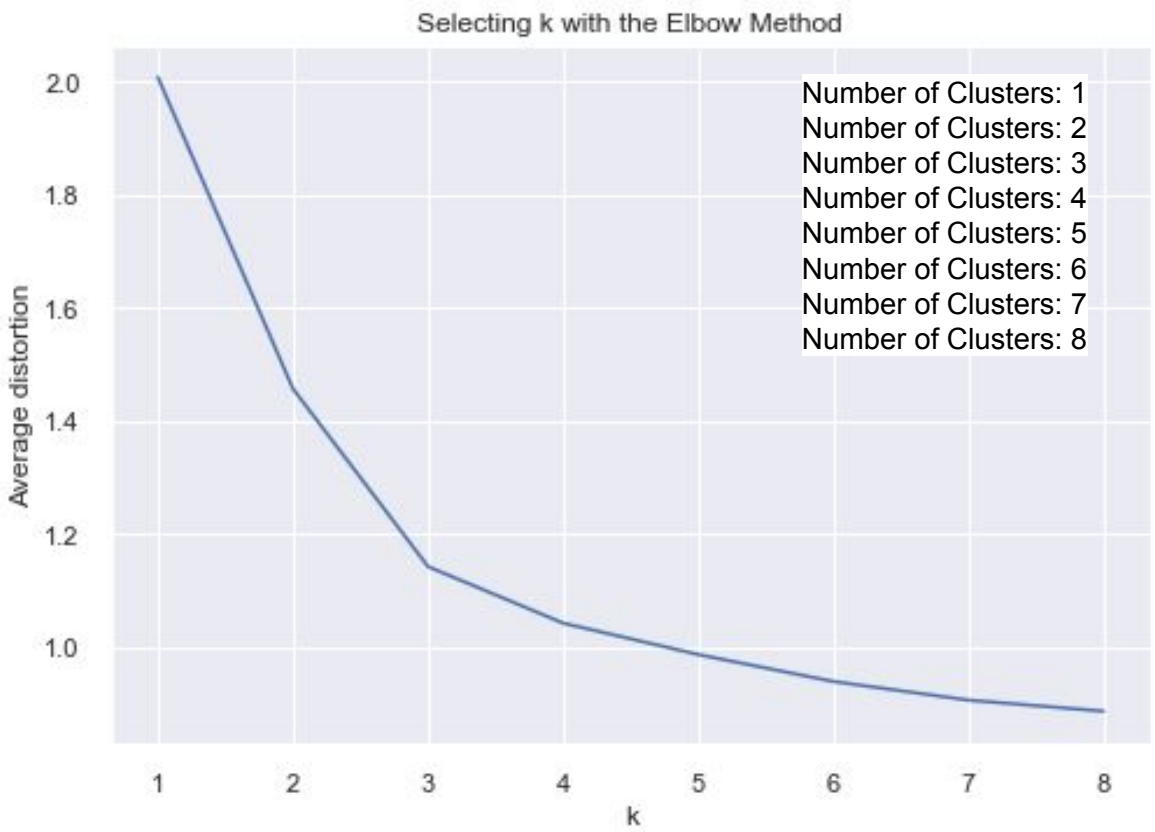
- It looks like those customers that have 5 or more Total_calls_made have an Avg_Credit_limit below 20000.
- Those customers that have 3 or 4 more Total_calls_made have an Avg_Credit_limit below 75000.
- Customers that have Total_calls_made 2 and below, have an Avg_Credit_Limit 200000 and below.



EDA - **Key Meaningful Observations**

- It looks like there is exclusive relationship between Total_visits_bank vs Total_visits_online vs Total_calls_made. Where customers that are predominant in one variable are weak in the others.
- Avg_Credit_limit seems to be divided in groups of those that have 75000 and below and those that have 75000 and above. This variable has the most outliers.

Applying K-means clustering algorithms



Average Distortion: 2.007896349270688
Average Distortion: 1.4576197022077821
Average Distortion: 1.1434401208195095
Average Distortion: 1.0435538595477063
Average Distortion: 0.9880591433704322
Average Distortion: 0.9404952836425913
Average Distortion: 0.9075861543551181
Average Distortion: 0.887984835729803

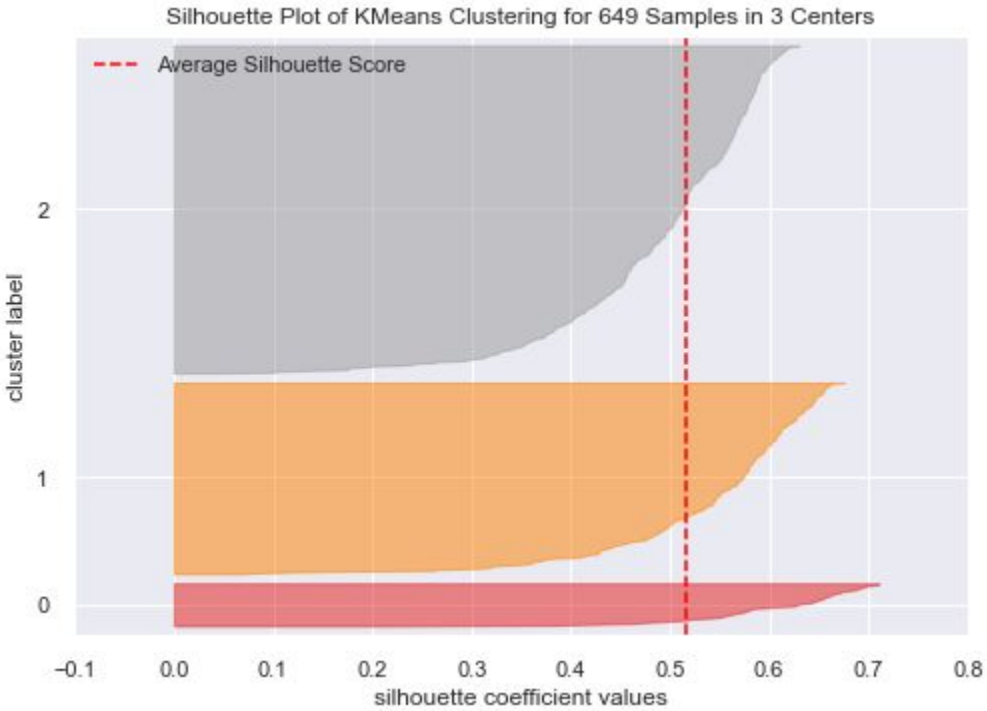
- Appropriate k seems to be 3 or 4.

Silhouette Score

For n_clusters = 2, silhouette score is 0.41800025566689647)
For n_clusters = 3, silhouette score is 0.516281010855363)
For n_clusters = 4, silhouette score is 0.3570238219413198)
For n_clusters = 5, silhouette score is 0.2722848313346344)
For n_clusters = 6, silhouette score is 0.25696498143767876)
For n_clusters = 7, silhouette score is 0.24796181778236623)
For n_clusters = 8, silhouette score is 0.22660108820428626)
For n_clusters = 9, silhouette score is 0.22077645663369874)

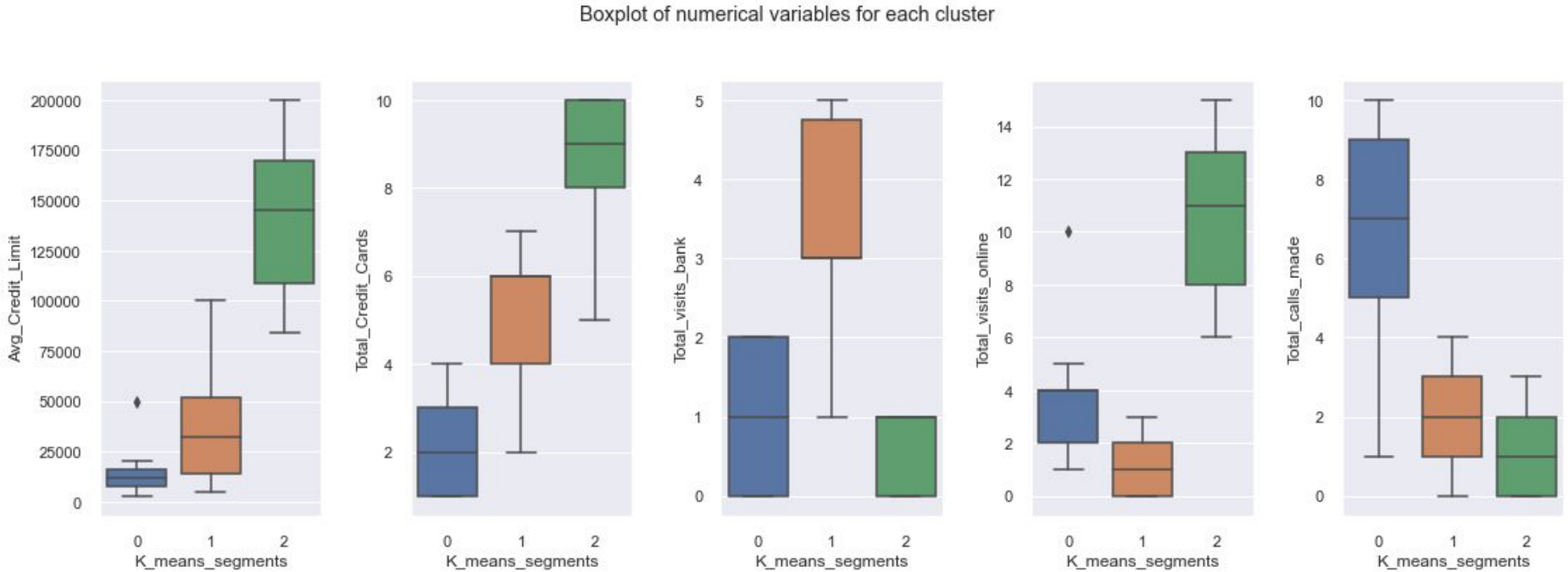
- Silhouette score for 3 is high (0.5162), so we will choose 3 as value of k.

Appropriate number of cluster with silhouette coefficients



- None of the clusters are overfitting.
- All clusters seem to be at the same distince from the mean value.

Insights - K-Means clustering

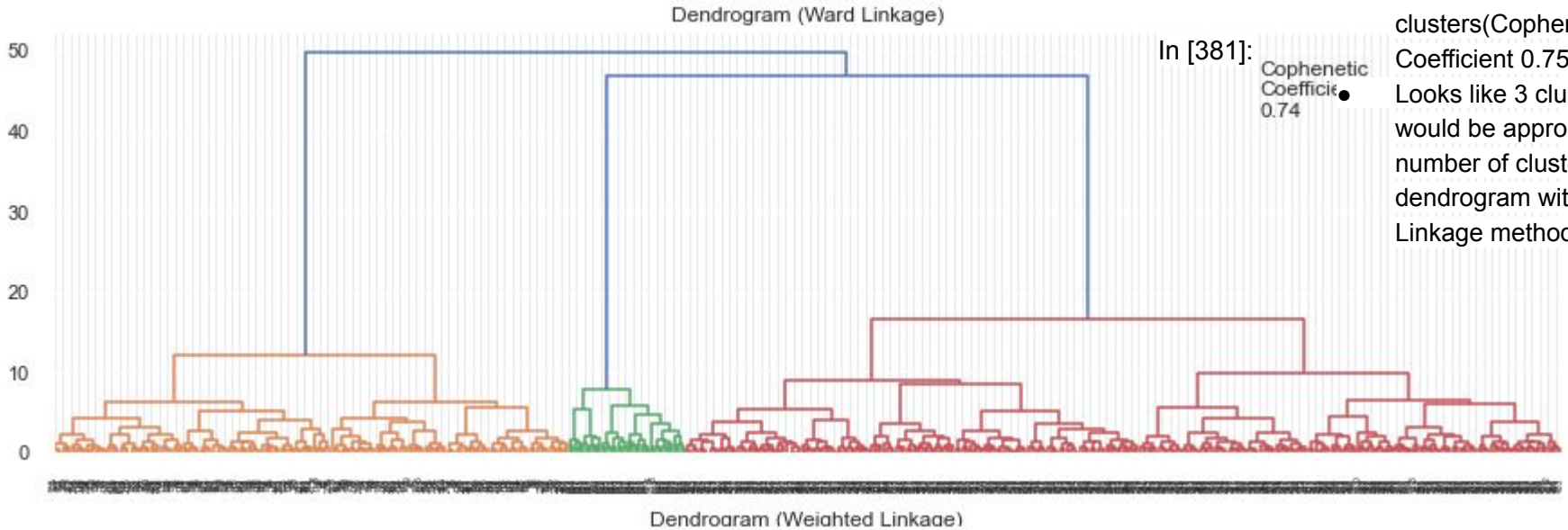


Insights - K-Means clustering

- **Cluster 0** :
  - This cluster contains customers with low Avg_Credit_Limit (less than 25000)
  - Also, low Total_Credit_Cards (3 or less)
  - This cluster prefers to interact with the bank via phone, hence the Total_calls_made is high.
  - This cluster does not prefers to interact with the bank via bank visits, nor online.
- **Cluster 1**:
  - This cluster contains customers with average Avg_Credit_Limit (between 20000 and 60000)
  - Also, average Total_Credit_Cards (between 4 and 6)
  - This cluster prefers to interact with the bank via bank visits, hence the Total_visit_bank is high.
  - This cluster does not prefers to interact with the bank via phone, nor online.
- **Cluster 2**:
  - This cluster contains customers with high Avg_Credit_Limit (between 100000 and 175000)
  - Also, high Total_Credit_Cards (between 8 and 10)
  - This cluster prefers to interact with the bank via online, hence the Total_calls_made is high.
  - This cluster does not prefers to interact with the bank via bank visits, nor phone.

Applying Hierarchical clustering

Cophenetic correlation for distance metrics euclidean and linkage method single is 0.7395135051413775
Cophenetic correlation for distance metrics euclidean and linkage method average is 0.8974425535306298
Cophenetic correlation for distance metrics euclidean and linkage method complete is 0.8794736468795109
Cophenetic correlation for distance metrics euclidean and linkage method centroid is 0.894471288720818
Cophenetic correlation for distance metrics euclidean and linkage method ward is 0.7425813590948763
Cophenetic correlation for distance metrics euclidean and linkage method weighted is 0.8551098644586315

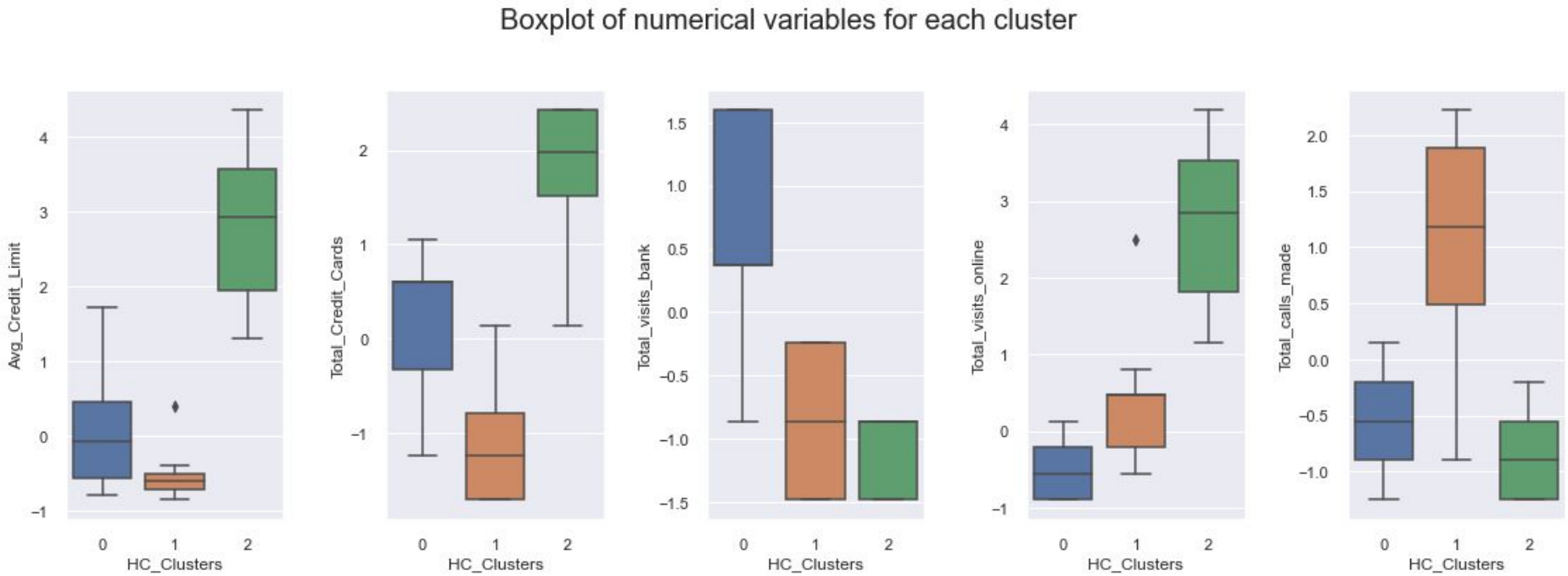


- Out of all the dendrogram we see, it is clear that dendrogram with ward linkage method gave us separate and distinct clusters(Cophenetic Coefficient 0.7576).
- Looks like 3 clusters would be appropriate number of cluster from dendrogram with Ward Linkage method

3 Clusters

In cluster 0 Total_Credit_Cards are: [2 7 5 4 6] In cluster 1 Total_Credit_Cards are: [3 2 4 1 5] In cluster 2 Total_Credit_Cards are: [6 5 9 8 10 7]	In cluster 0 Total_visits_bank are: [1 2 5 3 4] In cluster 1 Total_visits_bank are: [0 2 1] In cluster 2 Total_visits_bank are: [0 1]
In cluster 0 Total_visits_online are: [1 3 0 2] In cluster 1 Total_visits_online are: [10 1 2 5 4 3] In cluster 2 Total_visits_online are: [12 11 14 7 10 13 15 6 8 9]	n cluster 0 Total_calls_made are: [0 4 2 3 1] In cluster 1 Total_calls_made are: [9 8 1 2 7 5 6 4 10] In cluster 2 Total_calls_made are: [3 2 1 0]

Cluster Profile



Insights Hierarchical clustering

- **Cluster 0**:
  - This cluster contains customers with average Avg_Credit_Limit (between -0.5 and 0.5)
  - Also, average Total_Credit_Cards (between -0.2 and 0.7)
  - This cluster prefers to interact with the bank via bank visits, hence the Total_visit_bank is high.
  - This cluster does not prefers to interact with the bank via phone, nor online.
- **Cluster 1** :
  - This cluster contains customers with low Avg_Credit_Limit (less than -0.5)
  - Also, low Total_Credit_Cards (-0.8 or less)
  - This cluster prefers to interact with the bank via phone, hence the Total_calls_made is high.
  - This cluster does not prefers to interact with the bank via bank visits, nor online.
- **Cluster 2**:
  - This cluster contains customers with high Avg_Credit_Limit (between 2 and 3.5)
  - Also, high Total_Credit_Cards (between 1.5 and 3)
  - This cluster prefers to interact with the bank via online, hence the Total_calls_made is high.
  - This cluster does not prefers to interact with the bank via bank visits, nor phone.

Compare cluster K-means clusters and Hierarchical clusters - Cluster profiling

From the analysis above we can conclude that there si not much difference between K-Means Clustering and Hierarchical Clustering. Bot methods provided 3 clusters with similar charchteristics. And both methods have produced the same results.

Basically, in both methods, we have segmented the dataset in 3 clusters with the following profiles:

* Cluster 0 : is for Tier-1 accounts where customers have low `Avg_Credit_limit` and less `Total_Credit_Cards`. This segment prefers to interact with the bank by phone.

* Cluster 1 : is for Tier-2 accounts where customers have average `Avg_Credit_limit` and average `Total_Credit_Cards`. This segment prefers to interact with the bank by visits.

* Cluster 2 : is for Tier-3 accounts where customers have high `Avg_Credit_limit` and high `Total_Credit_Cards`. This segment prefers to interact with the bank online.

Actionable Insights & Recommendations

- Cluster 0 consists of Tier-1 account. The company can focus on offering more cards to this segment and by correlation, this segment can increate the Avg_Credit_Limit. Also, the bank can focus on providing better phone service support for this segment and cater to the products that this segment prefers.
- Cluster 1 consists of Tier-2 accounts. The company can provide better on-site support personel, since this segment prefers to physically visit the bank for their needs.
- Cluster 2 consists of Tier-3 accounts. The company can provide better on-line support, since this segment prefers to interact with the bank online.

greatlearning
Power Ahead

Happy Learning !

Esteban Ordenes

Post Graduate Program in Data Science and Business Analytics

PGP-DSBA-UTA-Dec20-A